IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Kennedy Opoku Asare
October 23, 2024

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

Executive Summary

- We use exploratory data analysis with SQL and Data visualization.
  We build interactive dashboard with Maps and Chars
- We develop a Machine Learning Model to predict whether the first stage will land successfully.
- We can predict whether the first stage will land successfully with 83% accuracy.
- More successful landings has been observed in later years with mores successes from CCAFS SLC 40 and KSC LC 39A.
- Missions targeting ES-L1, GEO, HEO, and SSO orbits are more likely to have successful landings, while those targeting GTO and SO are less successful.

Introduction

- SpaceX offers Falcon 9 launches at a significantly lower cost of 62 million dollars compared to other providers who charge upwards of 165 million dollars.
- This cost efficiency is largely due to the reusability of the Falcon 9's first stage, if they land successfully.
- In this project, we predict if the Falcon 9 first stage will land successfully, which directly impacts the cost of a launch.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - The SpaceX data was collected from Wikipedia and the SpaceX web API
- Perform data wrangling
  - Many steps were taken, including recoding bad outcomes such as False ASDS, 'False Ocean', 'False RTLS'
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

Two approaches were used to collected the data

Calling the SpaceX web API

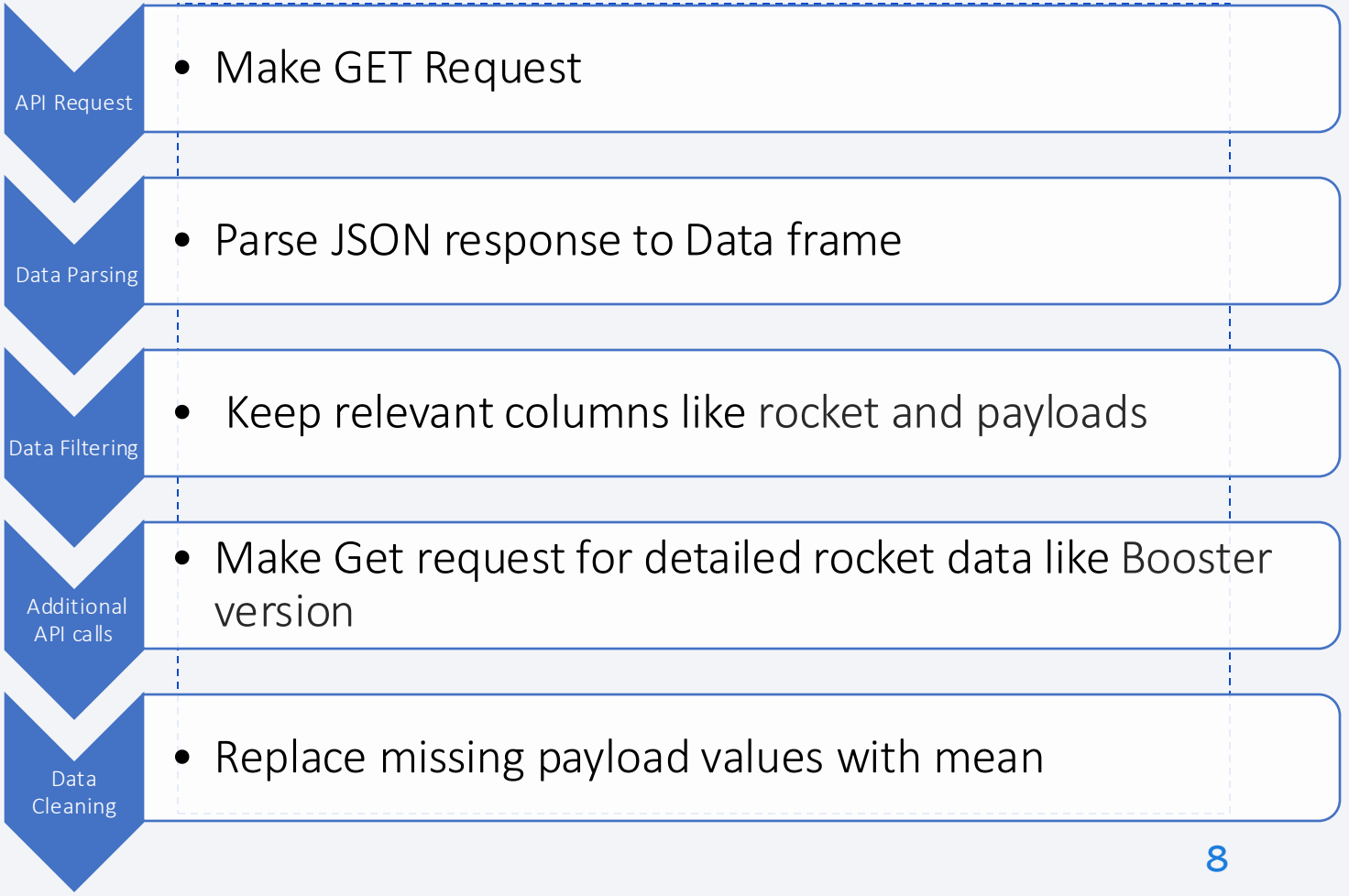
- Receiving the data in Json, normalizing the data.
- Data was filtered for specific columns, including rocket, payloads, launchpad and cores
- Additional API calls made for Booster version, payload mass and lauch site of rocket
- Missing values in PayloadMass are replaced with the column's mean.

Web Scrapping of the data from Wikipedia , using BeatifulSoup

- Read the web page with GET request
- Parse the web page with BeautifulSoup
- Scrap required data from tables
- Clean and structure the data
  - Setting required columns
- Save data with a dataframe

Data Collection – SpaceX API

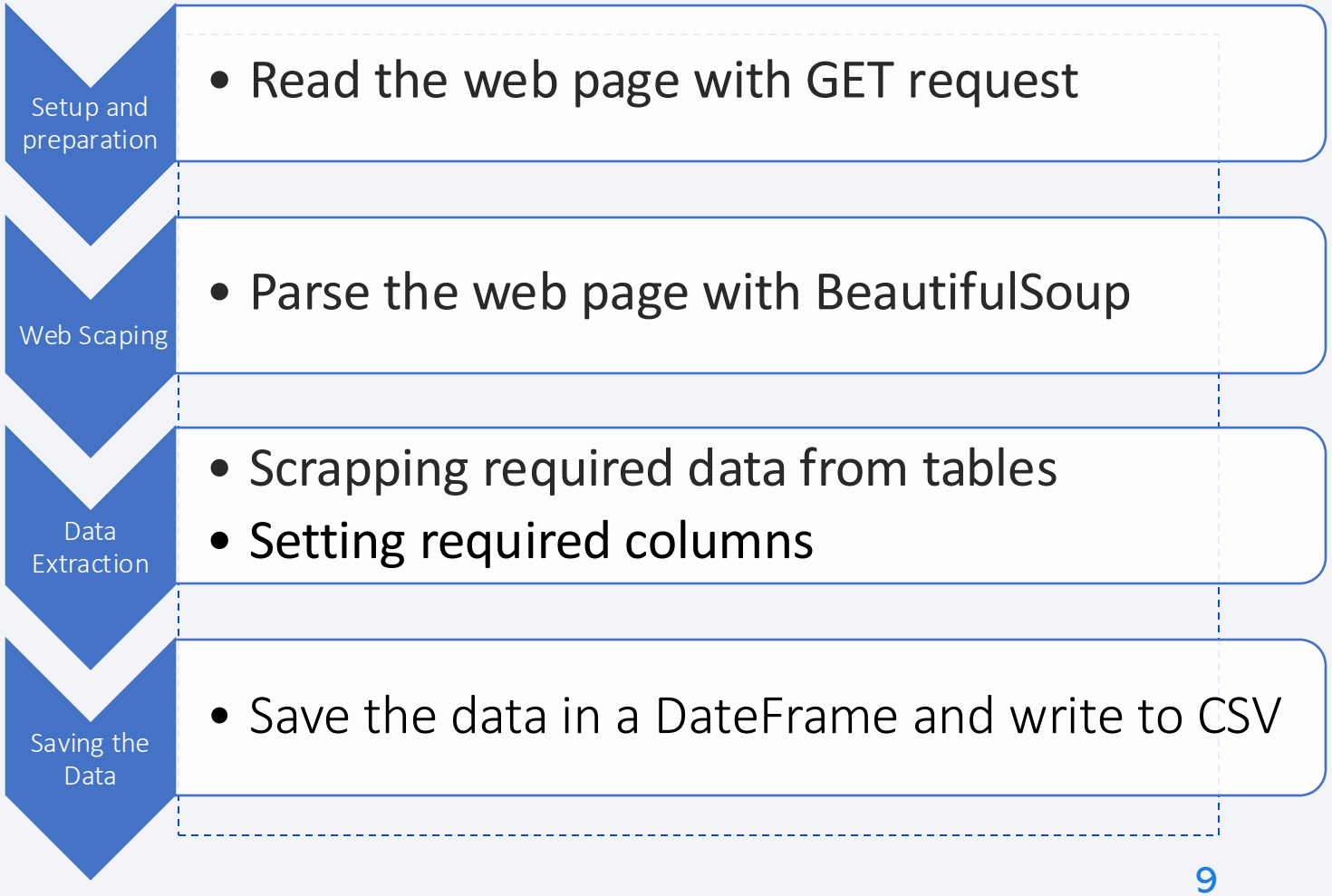
- GitHub URL https://github.com/kennedyopokuasare/IBM_datascience/blob/main/captone_project/spacex-data-collection-api.ipynb



Data Collection - Scraping

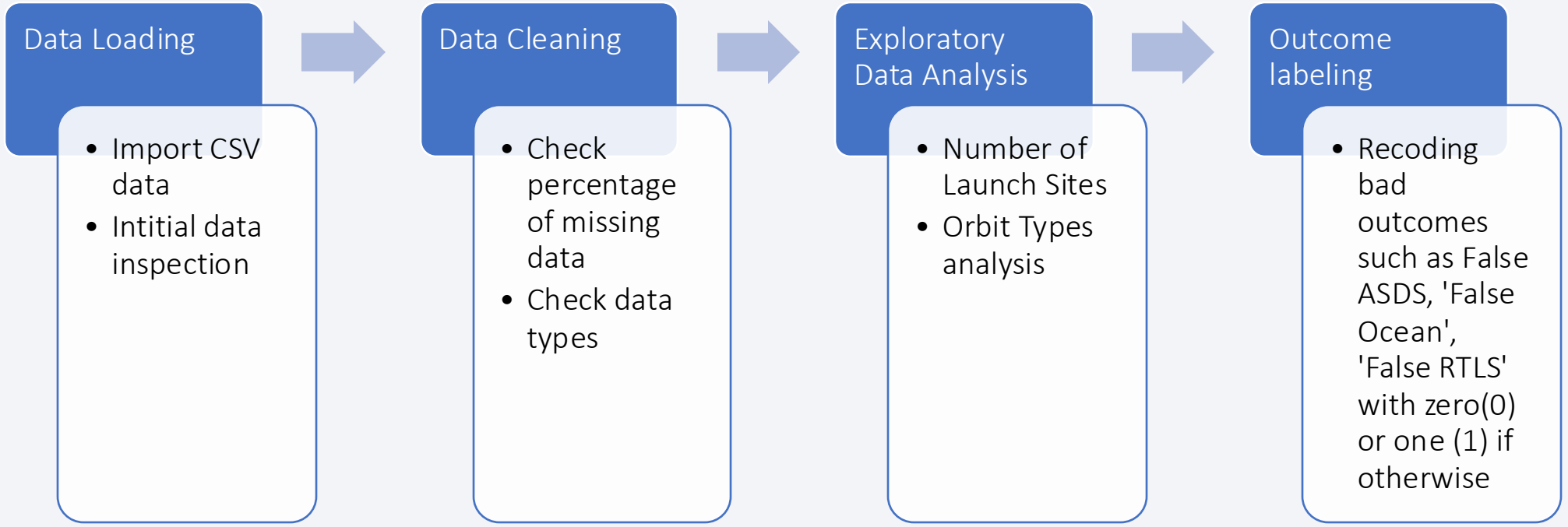
- GitHub URL

https://github.com/kennedyopokuasare/IBM_datascience/blob/main/captone_project/spacex-data-webscraping.ipynb



Data Wrangling

GitHub URL: https://github.com/kennedyopokuasare/IBM_datascience/blob/main/captone_project/spacex-data_wrangling.ipynb



EDA with Data Visualization

GitHub URL: https://github.com/kennedyopokuasare/IBM_datascience/blob/main/captone_project/spacex-eda-dataviz.ipynb

- Flight Number vs. Payload Mass
  - Category plot to observe how the FlightNumber and PayloadMass affect the launch outcome.
- Flight Number vs. Launch Site
  - Category plot to visualize the relationship between FlightNumber and LaunchSite.
- Payload Mass vs. Launch Site
  - Scatter plot to explore the relationship between PayloadMass and LaunchSite.
- Success Rate by Orbit Type
  - Bar chat to visualize the success rate of each orbit type
- Flight Number vs. Orbit Type
  - Scatter plot to examine the relationship between FlightNumber and Orbit.
- Payload Mass vs. Orbit Type
  - To reveal the relationship between PayloadMass and Orbit
- Yearly Success Rate Trend (Line Chart)
  - To plot the average success rate over the years

EDA with SQL

GitHub URL: https://github.com/kennedyopokuasare/IBM_datascience/blob/main/captone_project/spacex-eda-sql-coursera_sqllite.ipynb

- **Identify Unique Launch Sites:** To find all distinct launch sites used in the space missions.
- **Filter Launch Sites by Prefix:** To retrieve records of launches from sites starting with 'CCA', limited to 5 entries.
- **Calculate Total Payload for NASA (CRS):** To determine the total payload mass carried by boosters for NASA's CRS missions.
- **Compute Average Payload for Specific Booster:** To find the average payload mass for missions using the F9 v1.1 booster version.
- **Find First Successful Ground Pad Landing:** To list the date of the first successful landing on a ground pad.
- **Identify Successful Drone Ship Landings:** To list booster versions that successfully landed on a drone ship with payloads between 4000 and 6000 kg.
- **Count Mission Outcomes:** To tally the number of successful and failed mission outcomes.
- **Identify Boosters with Maximum Payload:** To find booster versions that carried the maximum payload mass using a subquery.
- **List Failures in 2015:** To display records of failed drone ship landings, booster versions, and launch sites for the year 2015.
- **Rank Landing Outcomes by Frequency:** To rank the count of different landing outcomes between specific dates in descending order.

Build an Interactive Map with Folium

GitHub URL: https://github.com/kennedyopokuasare/IBM_datascience/blob/main/captone_project/spacex-map-viz.ipynb

- **Markers:** To pinpoint specific locations on the map, such as launch sites and points of interest.
- **Circles:** To highlight areas around specific points, providing a visual cue for proximity or influence.
- **Polylines:** To draw lines between two points, illustrating connections or distances.
- **Mouse Position:** To dynamically display the coordinates of the mouse pointer on the map, aiding in identifying precise locations.

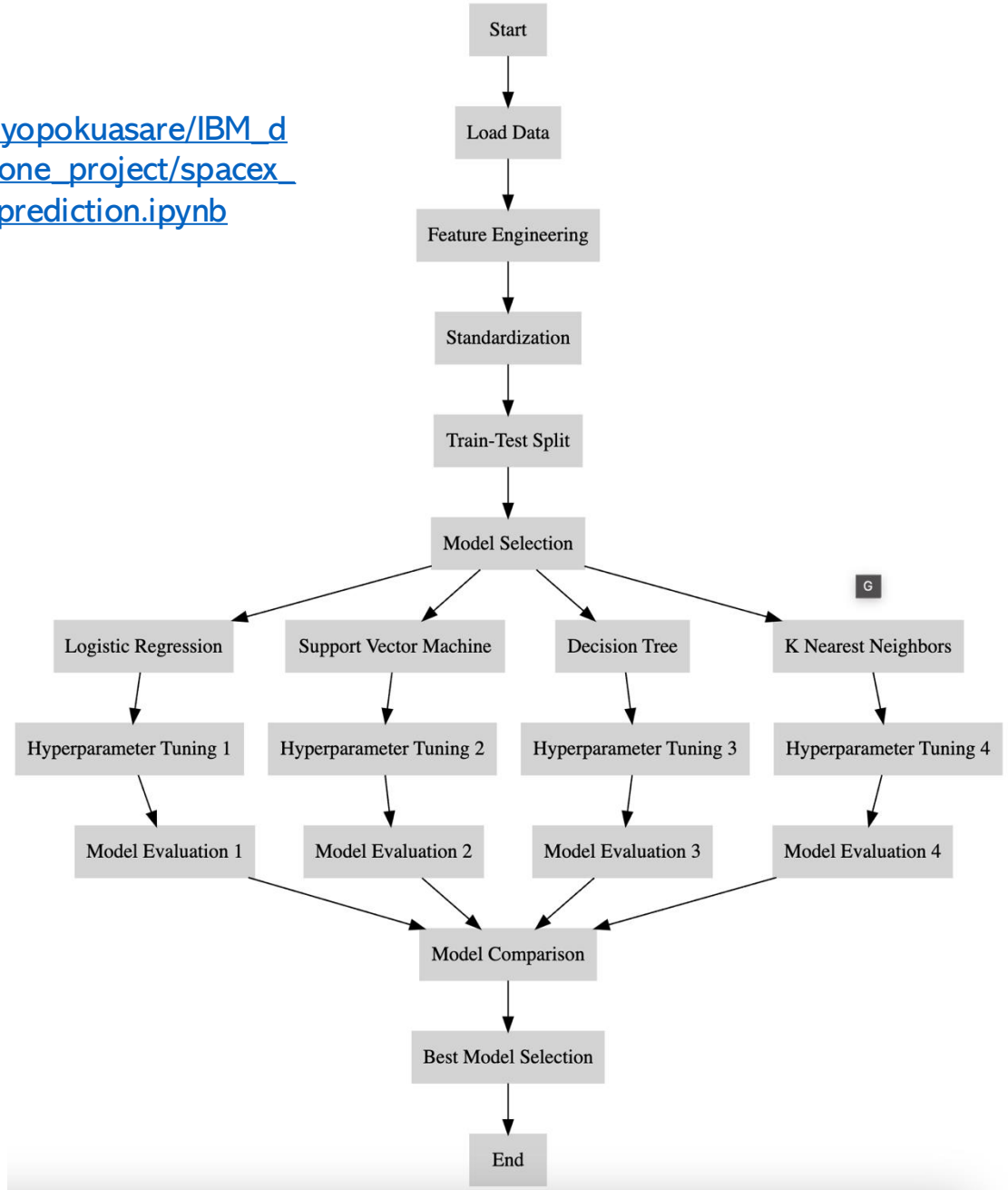
Build a Dashboard with Plotly Dash

GitHub URL: https://github.com/kennedyopokuasare/IBM_datascience/blob/main/captone_project/spacex_dash_app.py

- **Dropdown for Launch Site Selection:** Allows users to select a specific launch site or view data for all sites to filter the data displayed in the plots.
- **Pie Chart for Launch Success:** Visualizes the total number of successful launches for all sites or the success vs. failure counts for a selected site to provide a quick overview of launch outcomes.
- **Range Slider for Payload Selection:** Allows users to filter the data based on a range of payload masses to explore how payload mass affects launch success.
- **Scatter Chart for Payload vs. Launch Success:** Shows the correlation between payload mass and launch success to help identify patterns and correlations between payload mass and launch outcomes.

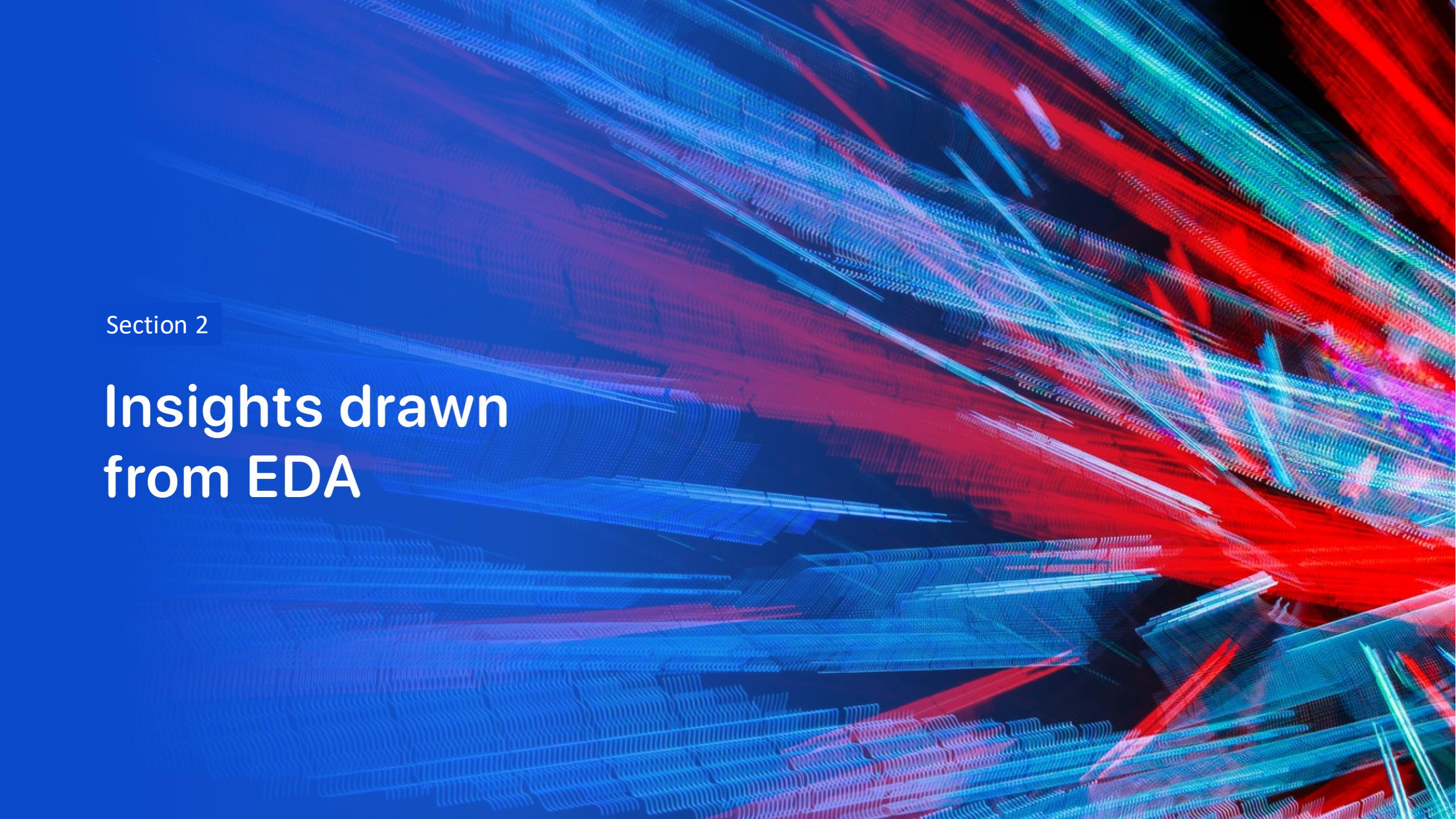
Predictive Analysis (Classification)

GitHub URL:
https://github.com/kennedyopokuasare/IBM_datascience/blob/main/captone_project/spacex_machine%20learning%20prediction.ipynb

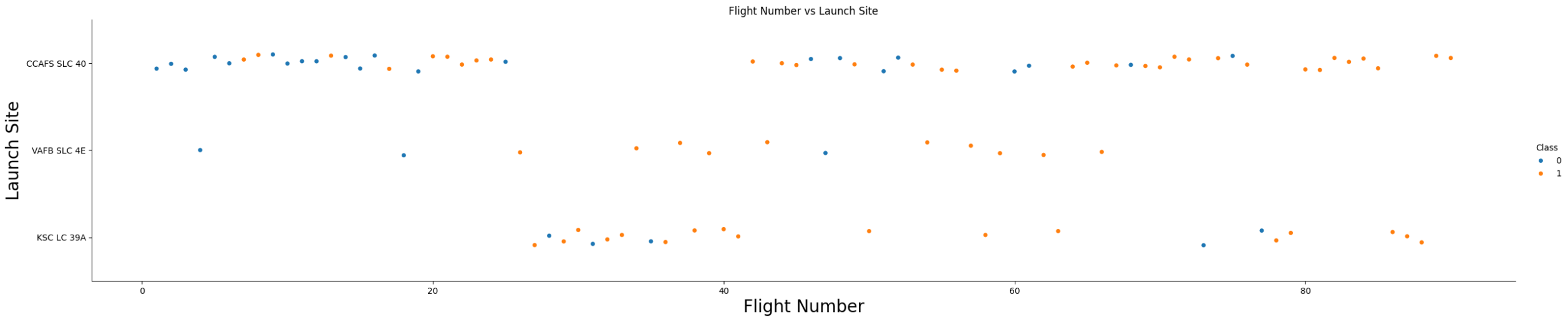


Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

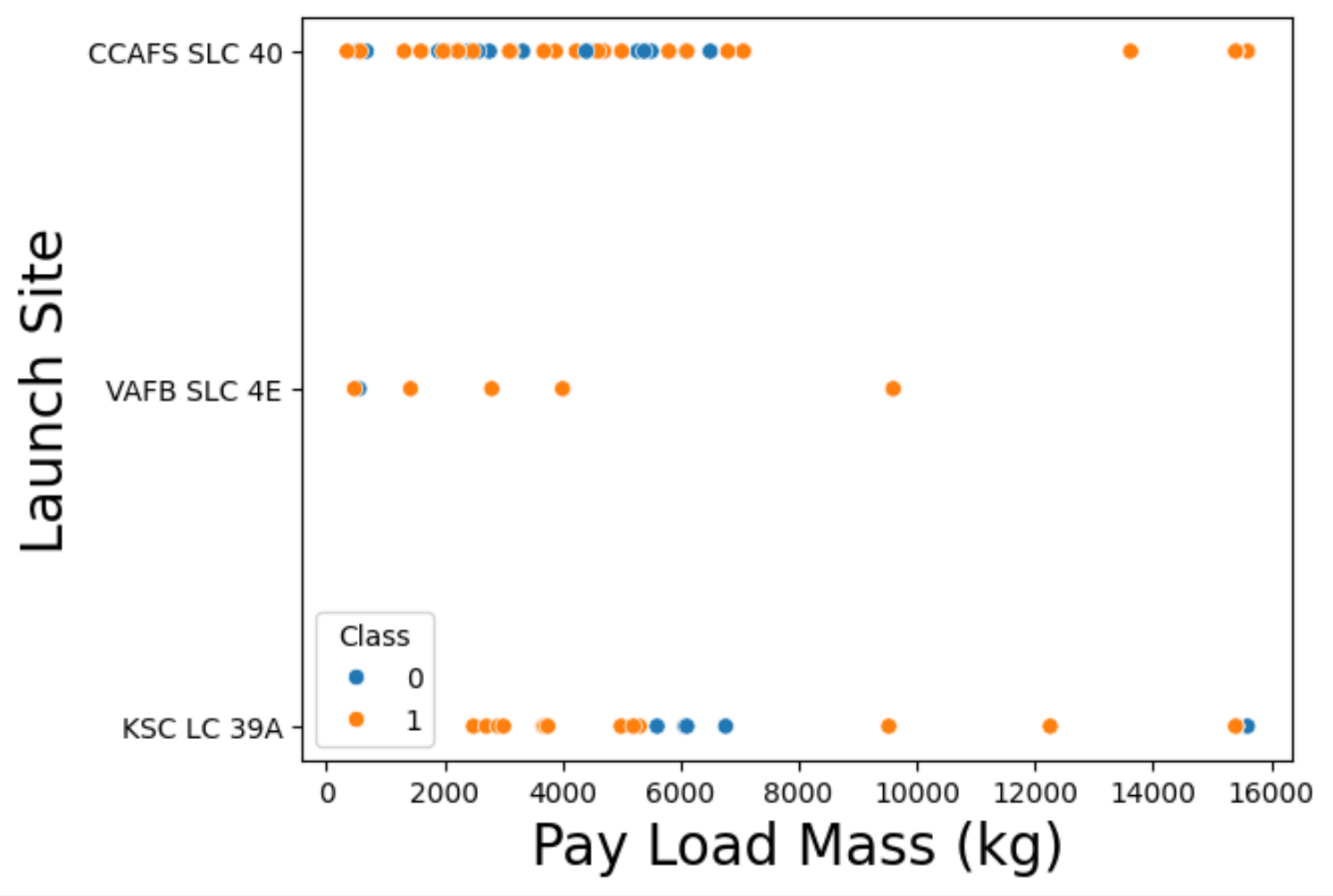

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



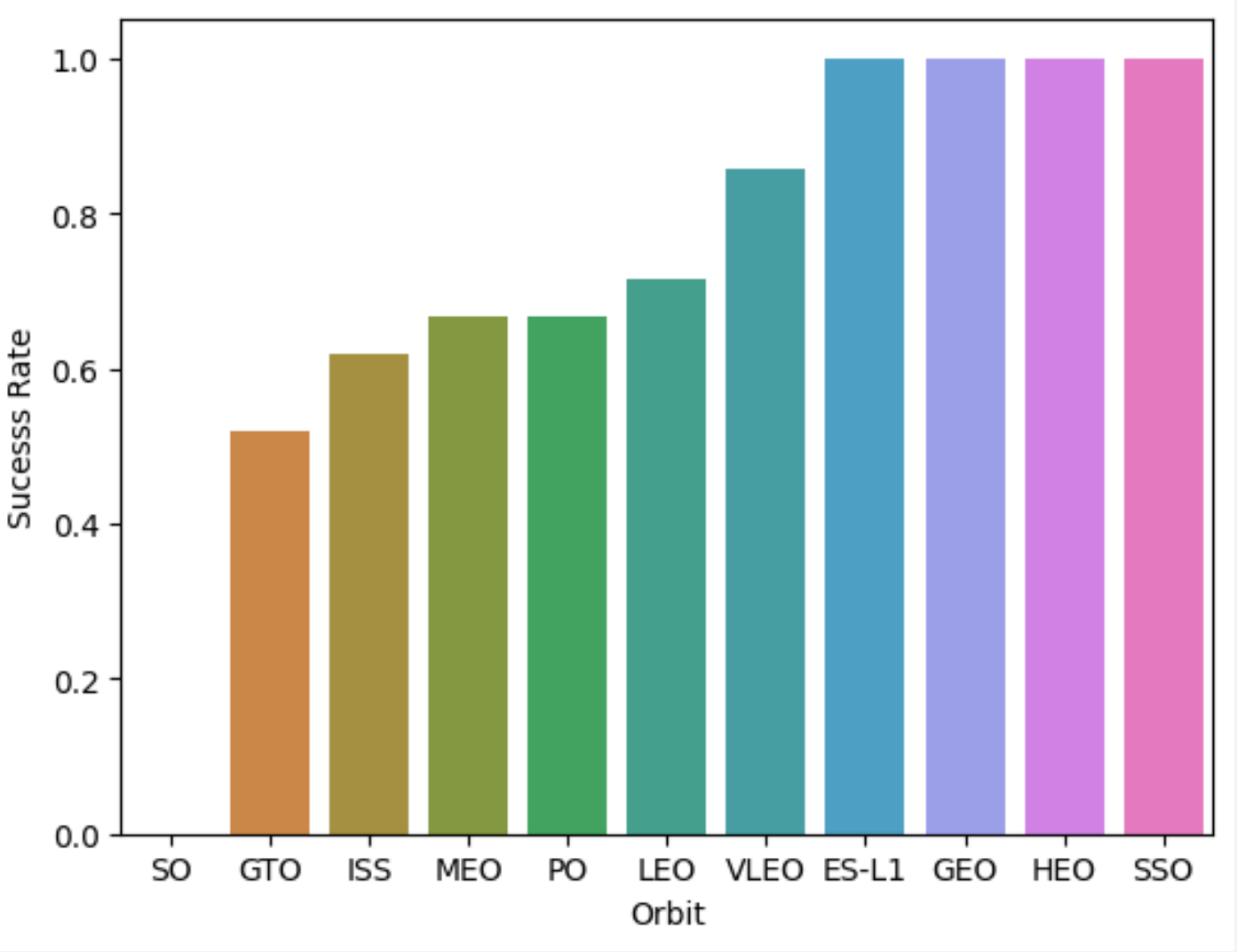
- Most launches occur from CCAFS SLC 40 and KSC LC 39A, with fewer from VAFB SLC 4E.
- As the flight number increases, indicating more recent launches, there is a trend towards more successful landings (Class 1), indicating an improvement in success rates over time
- Successful landings (orange dots) are more frequent in later flights, especially from CCAFS SLC 40 and KSC LC 39A.

Payload vs. Launch Site



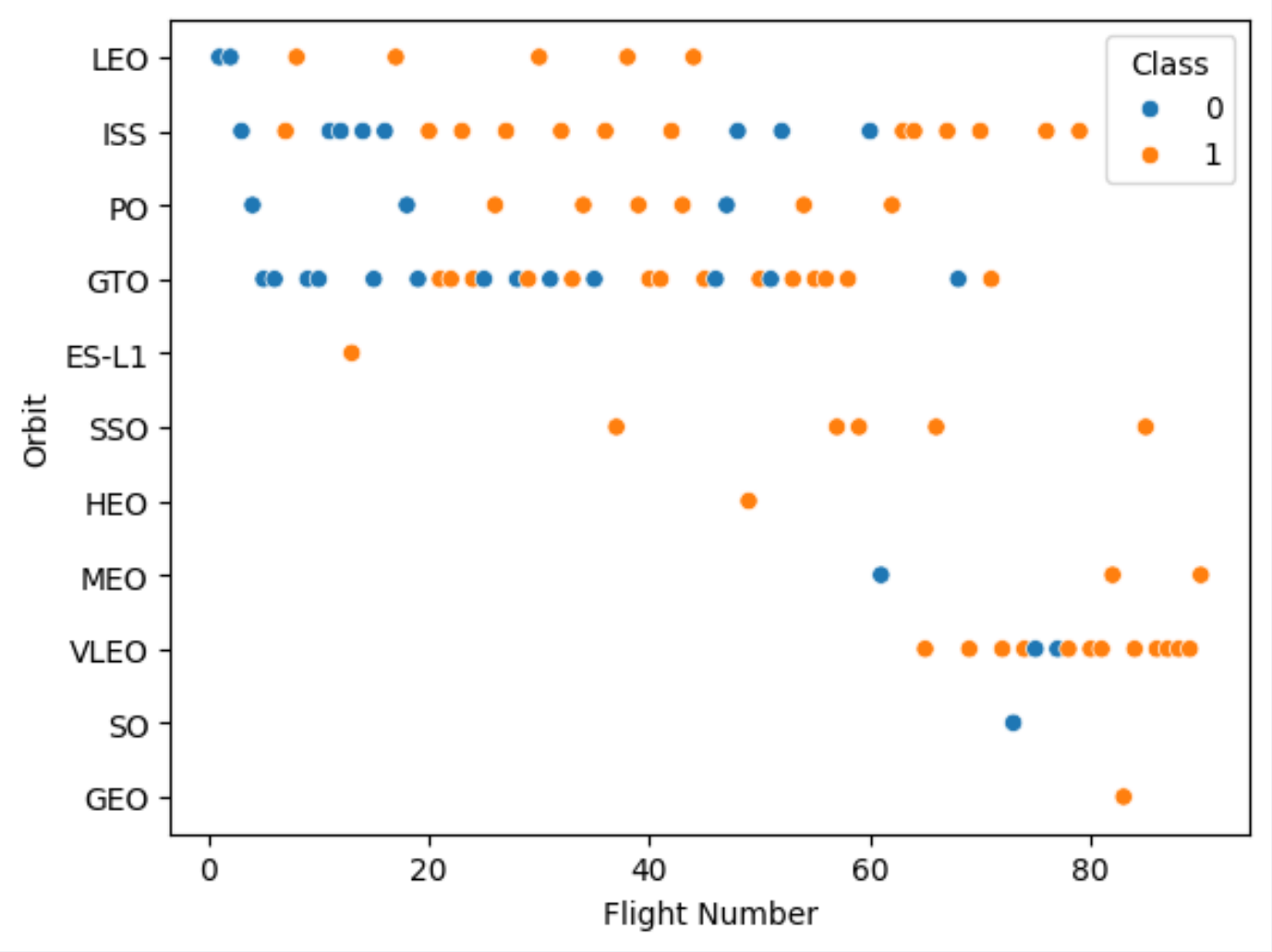
- CCAFS SLC 40 and KSC LC 39A have a wider range of payload masses compared to VAFB SLC 4E.
- VAFB SLC 4E has fewer launches and generally handles lower payload masses
- Successful landings (Class 1, orange dots) occur across all payload masses at CCAFS SLC 40 and KSC LC 39A.
- VAFB SLC 4E shows fewer data points, but successful landings are present, with no rockets launched for heavy payload mass (greater than 10000)

Success Rate vs. Orbit Type



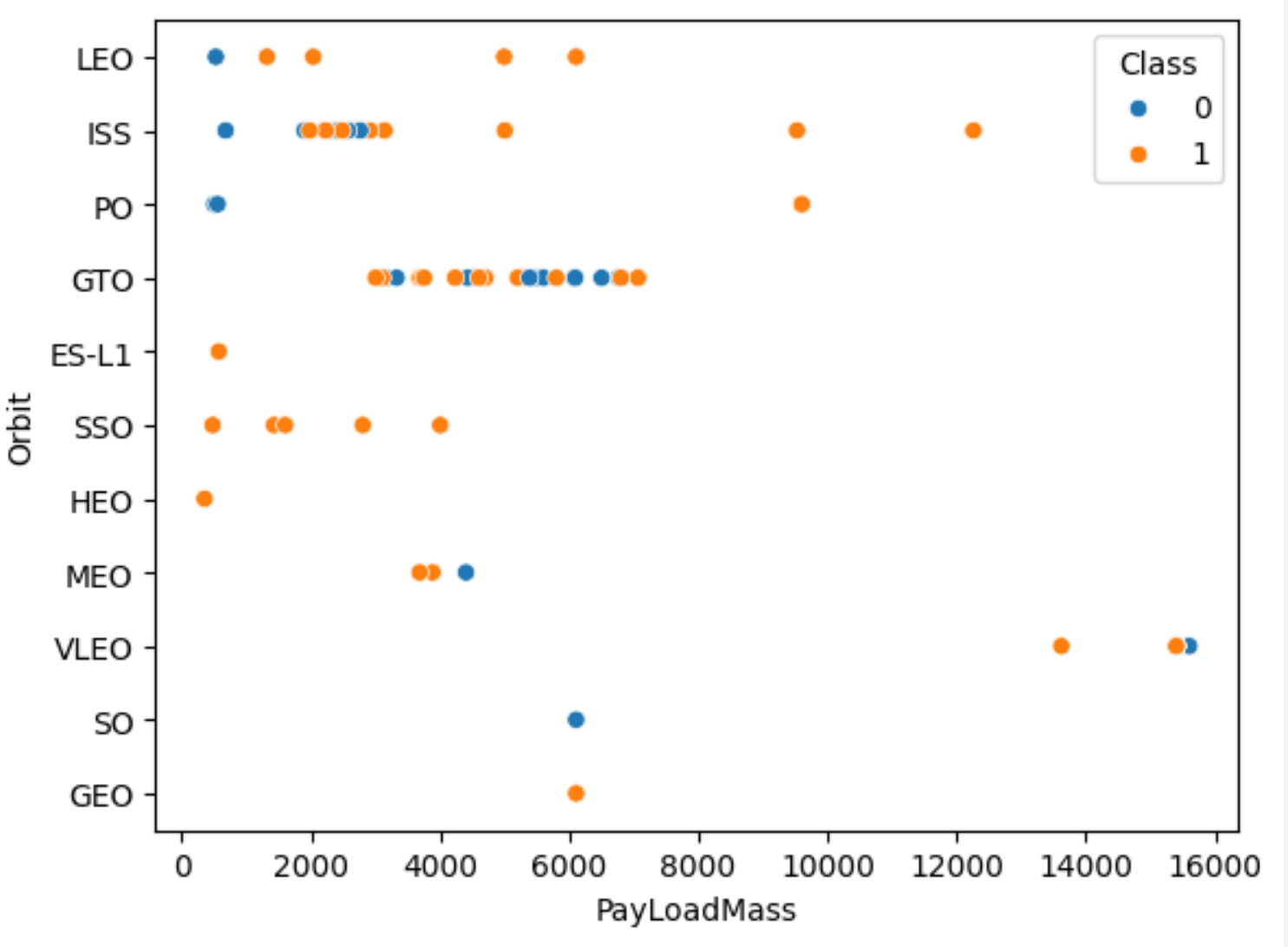
- ES-L1, GEO, HEO, and SSO orbits have the highest success rates, all reaching 100%.
- LEO, PO, MEO, and ISS have moderate success rates, ranging from about 60% to 80%
- GTO and SO have the lowest success rates, with SO being the lowest.
- This suggests that missions targeting ES-L1, GEO, HEO, and SSO orbits are more likely to have successful landings, while those targeting GTO and SO are less successful.

Flight Number vs. Orbit Type



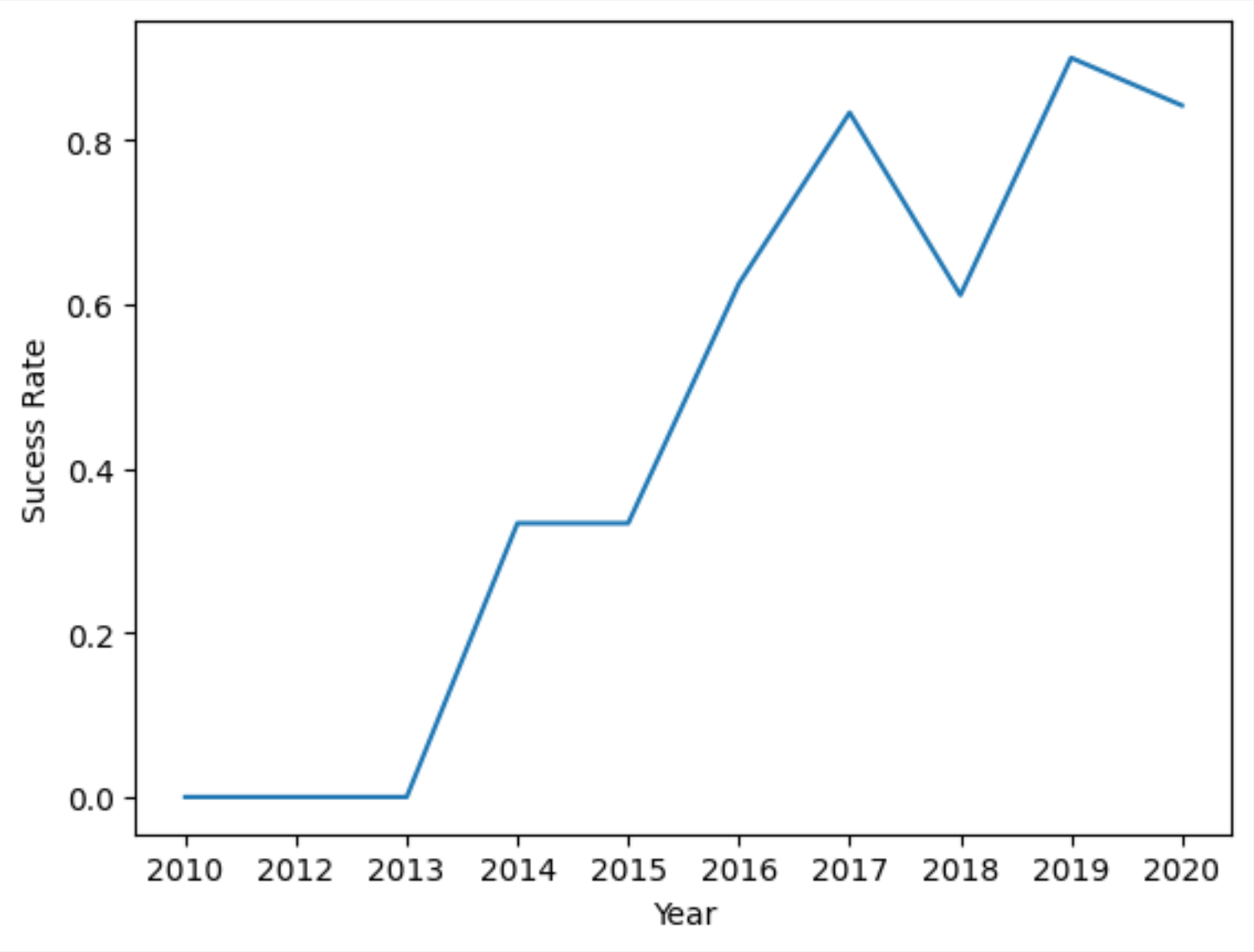
- As the flight number increases, indicating more recent launches, there is a trend towards more successful landings (Class 1, orange dots).
- Some orbits, like ES-L1, SSO, and GEO, show a higher concentration of successful landings.
- LEO and ISS have a mix of successful and unsuccessful landings, with a slight increase in success over time.
- GTO shows a more balanced mix of outcomes, indicating challenges in achieving consistent success.

Payload vs. Orbit Type



- GTO orbit shows a concentration of payloads around 6000 kg, with mixed success.
- LEO and ISS orbits handle a wide range of payload masses, with many successful landings (Class 1, orange dots).
- Higher payload masses, especially in VLEO and GEO, tend to have successful outcomes.

Launch Success Yearly Trend



- The yearly launch success rate since 2013 kept increasing till 2017, then there was a fluctuation in 2018
- From 2019 to 2020, the success rate rebounded to above 80%
- Overall, Lauch success rates has significantly improved over the years.

All Launch Site Names

- The unique launch sites in the SpaceX dataset are:
- **CCAFS SLC 40**: Cape Canaveral Air Force Station Space Launch Complex 40.
- **VAFB SLC 4E**: Vandenberg Air Force Base Space Launch Complex 4E.
- **KSC LC 39A**: Kennedy Space Center Launch Complex 39A.

CCAFS SLC 40 and CCAFS LC 40 are the same. This needs data cleaning to merge the two.

Launch Site Names Begin with 'CCA'

- The Launch Site beginning with "CCA," is CCAFS LC-40 (Cape Canaveral Air Force Station Launch Complex 40).
- All entries use the Falcon 9 v1.0 booster, indicating early missions in the Falcon 9 program.

Total Payload Mass

- The total payload mass carried by boosters launched by NASA (CRS) is 45596 kilograms

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 is 2534.67 Kilograms

First Successful Ground Landing Date

- The first date of successful landing outcome on ground pad was December 22, 2015

Successful Drone Ship Landing with Payload between 4000 and 6000

- The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 were

F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes were

Mission Outcome	Total Outcome
Failure	1
Success	100

- With data cleaning, all successes were merged to obtain 100 successes.

Boosters Carried Maximum Payload

- The names of the booster which have carried the maximum payload mass were

Booster Version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

All boosters with sub versions of F9 B5

2015 Launch Records

- The failed landing outcomes in drone ship, with their booster versions, and launch site names for in year 2015

Month	Landing Outcome	Booster Version	Launch Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- The launches occurred in January (01) and April (04).
- Both launches took place at CCAFS LC-40 (Cape Canaveral Air Force Station Launch Complex 40).
- The booster versions used were F9 v1.1 B1012 and F9 v1.1 B1015, both part of the Falcon 9 v1.1 series.

Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

- The Rank of landing outcomes between the date 2010-06-04 and 2017-03-20, are

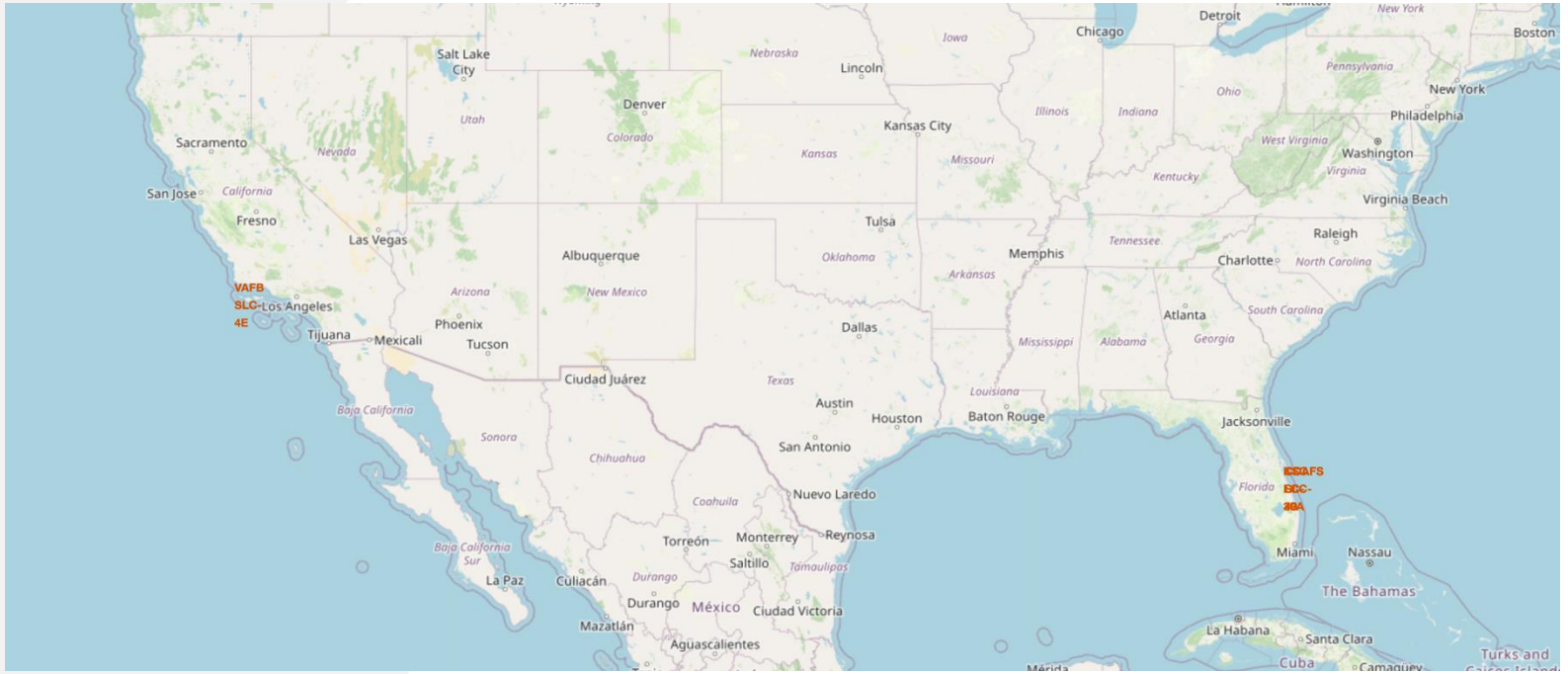
Landing Outcome	Total Outcomes
Controlled (ocean)	3
Failure (drone ship)	5
Failure (parachute)	2
No attempt	10
Precluded (drone ship)	1
Success (drone ship)	5
Success (ground pad)	3
Uncontrolled (ocean)	2

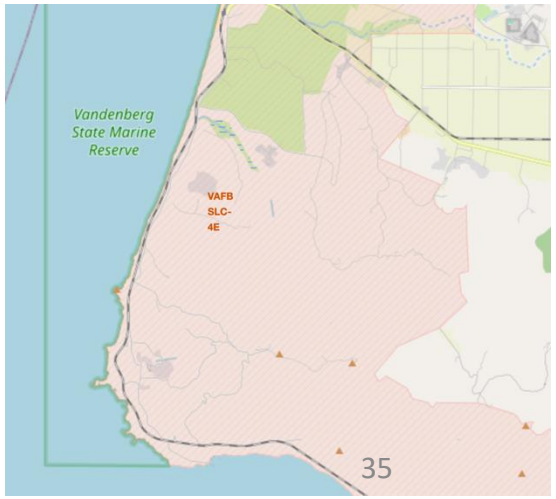
Section 3

Launch Sites Proximities Analysis

Launch Sites on the US Map

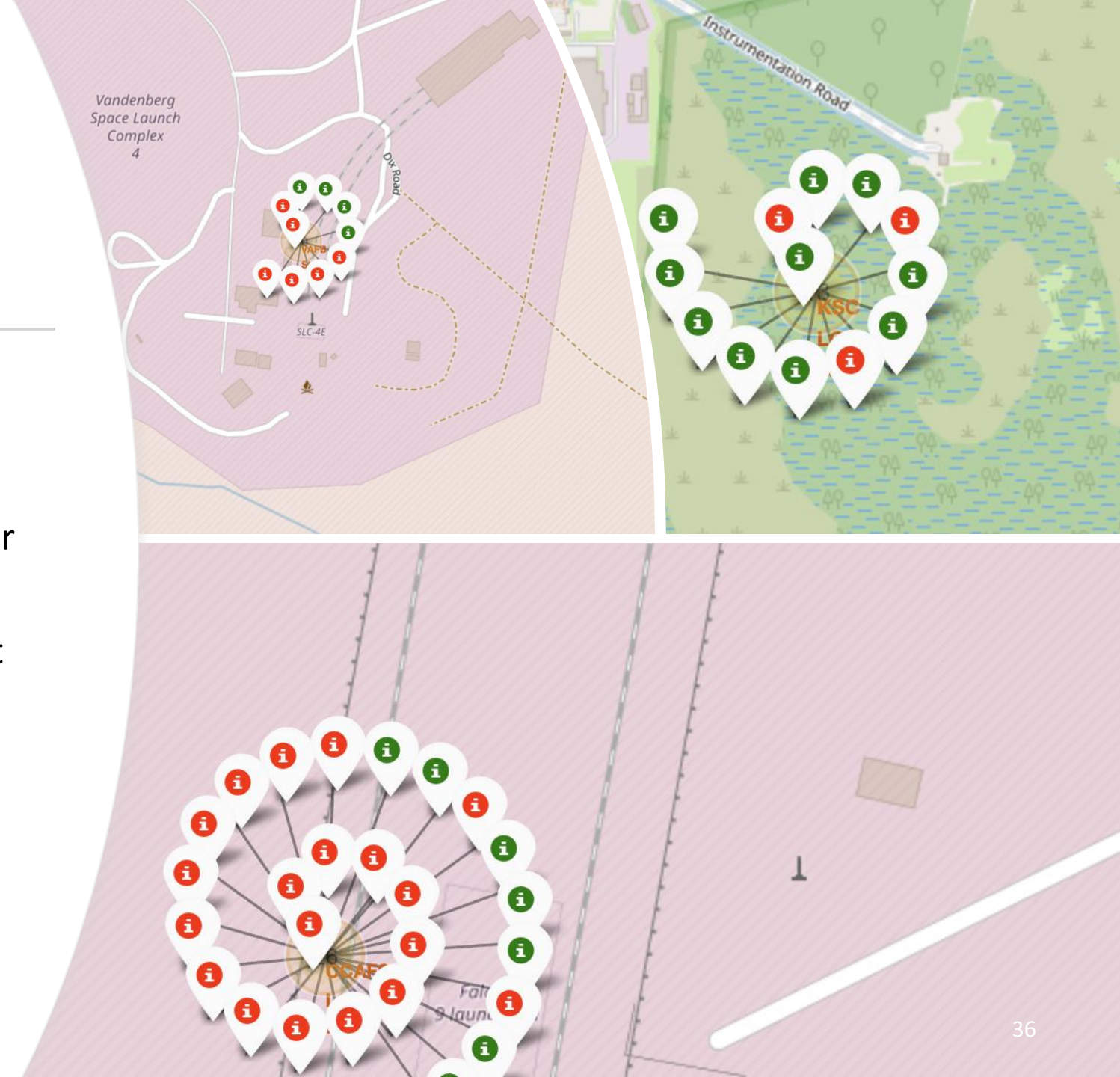
- Labeling of Launch sites on the east and west coasts of the USA





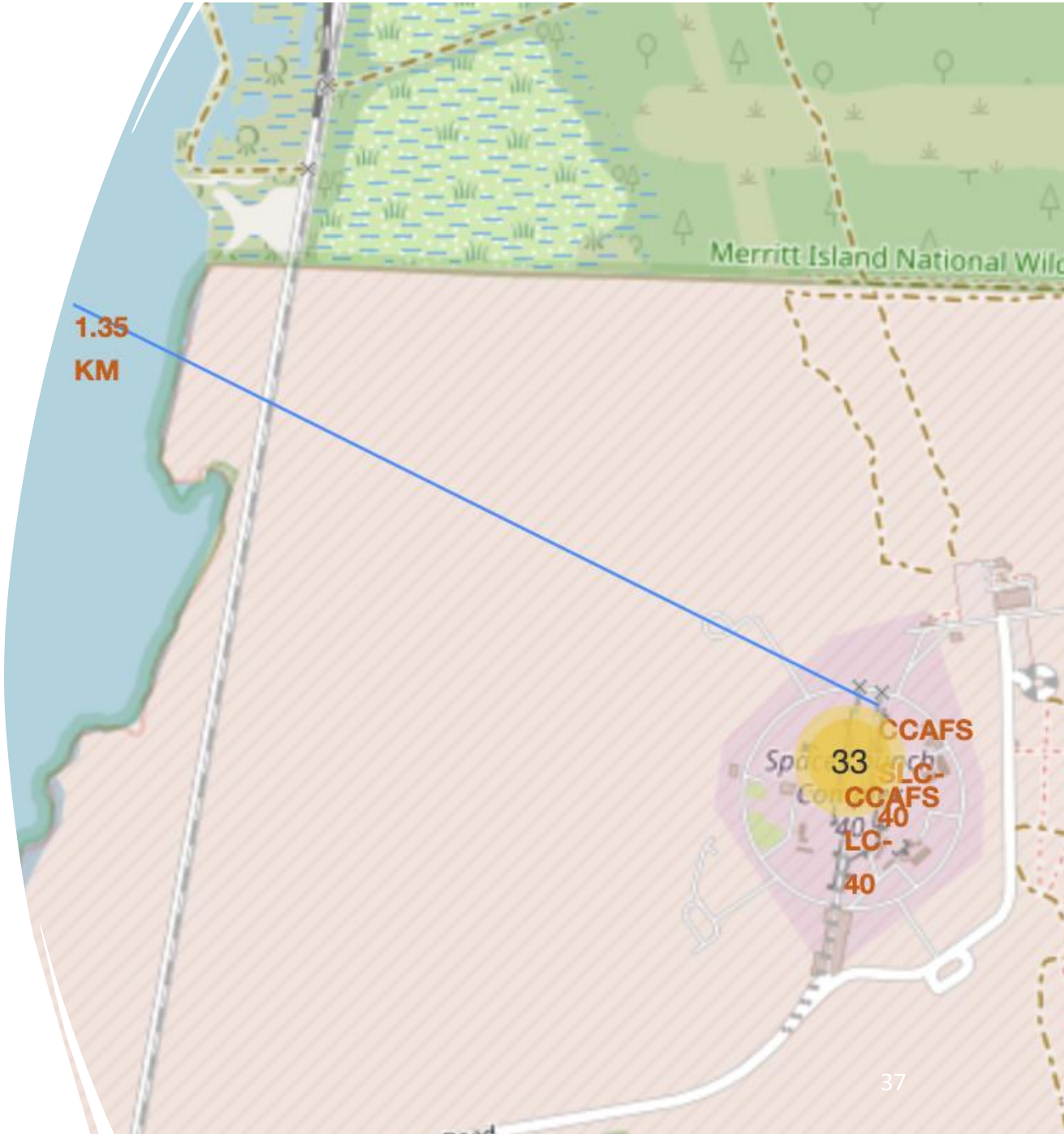
Launch site outcomes

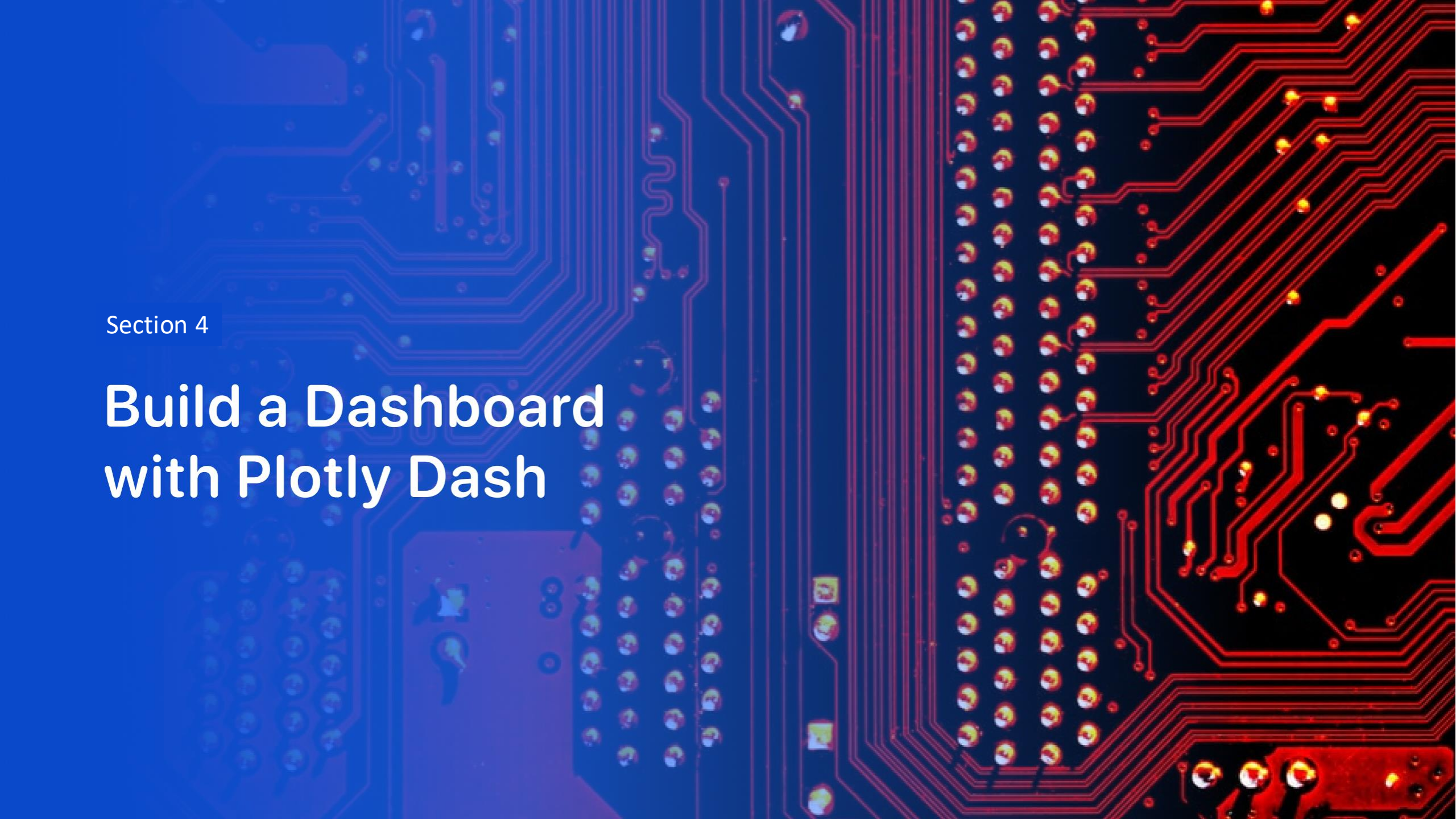
- Successful outcome at the various launch sites are labeled in green with a marker, or red for failure outcomes.
- This help identifies which specific spots at the Launch sites are more likely to have a good successful outcome, all other things being equal.



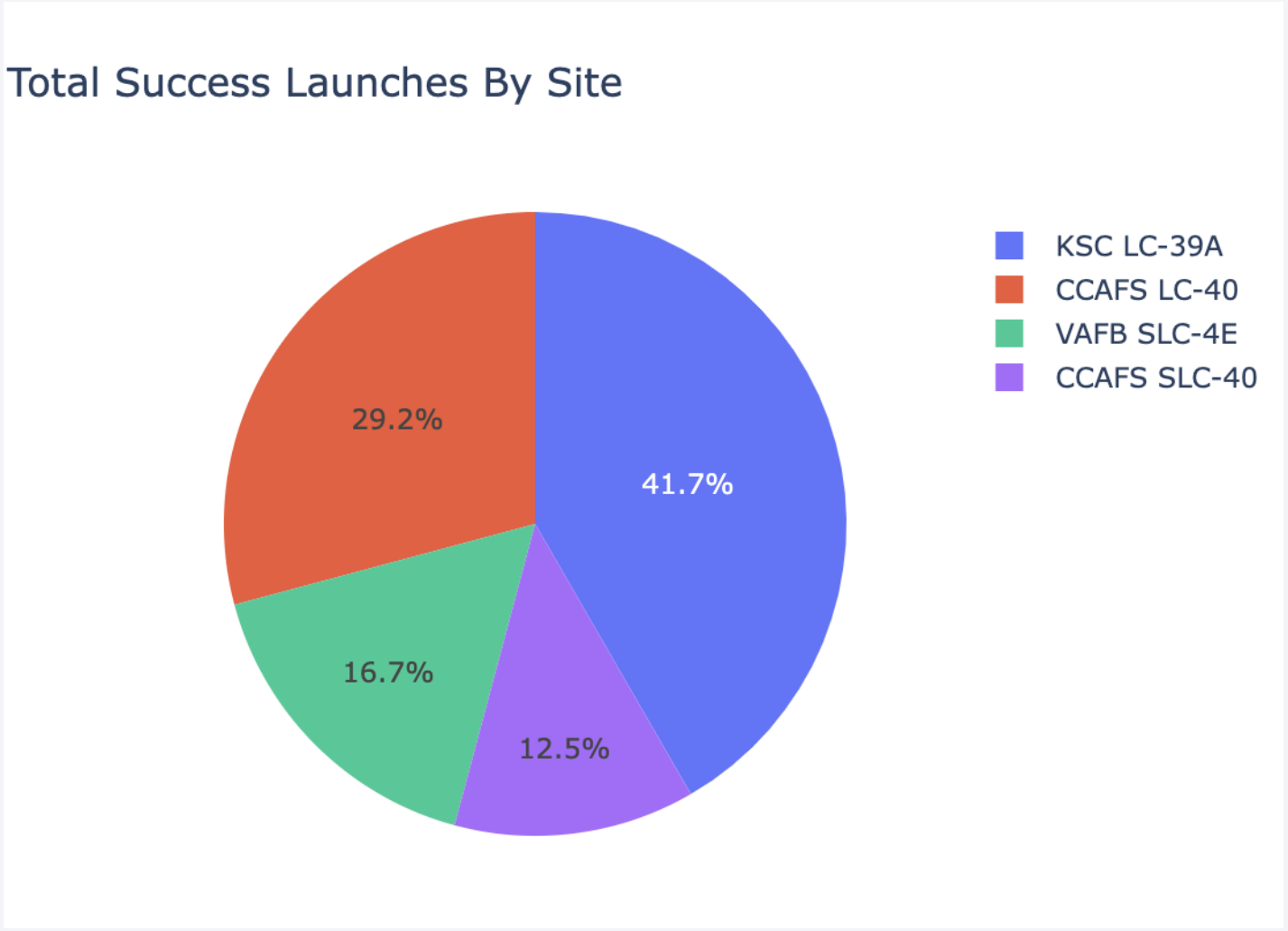
Launch Site proximity to Points of Interest

- A polyline drawn from the launch site to coastal area showing a distance of 1.35 Kilometers



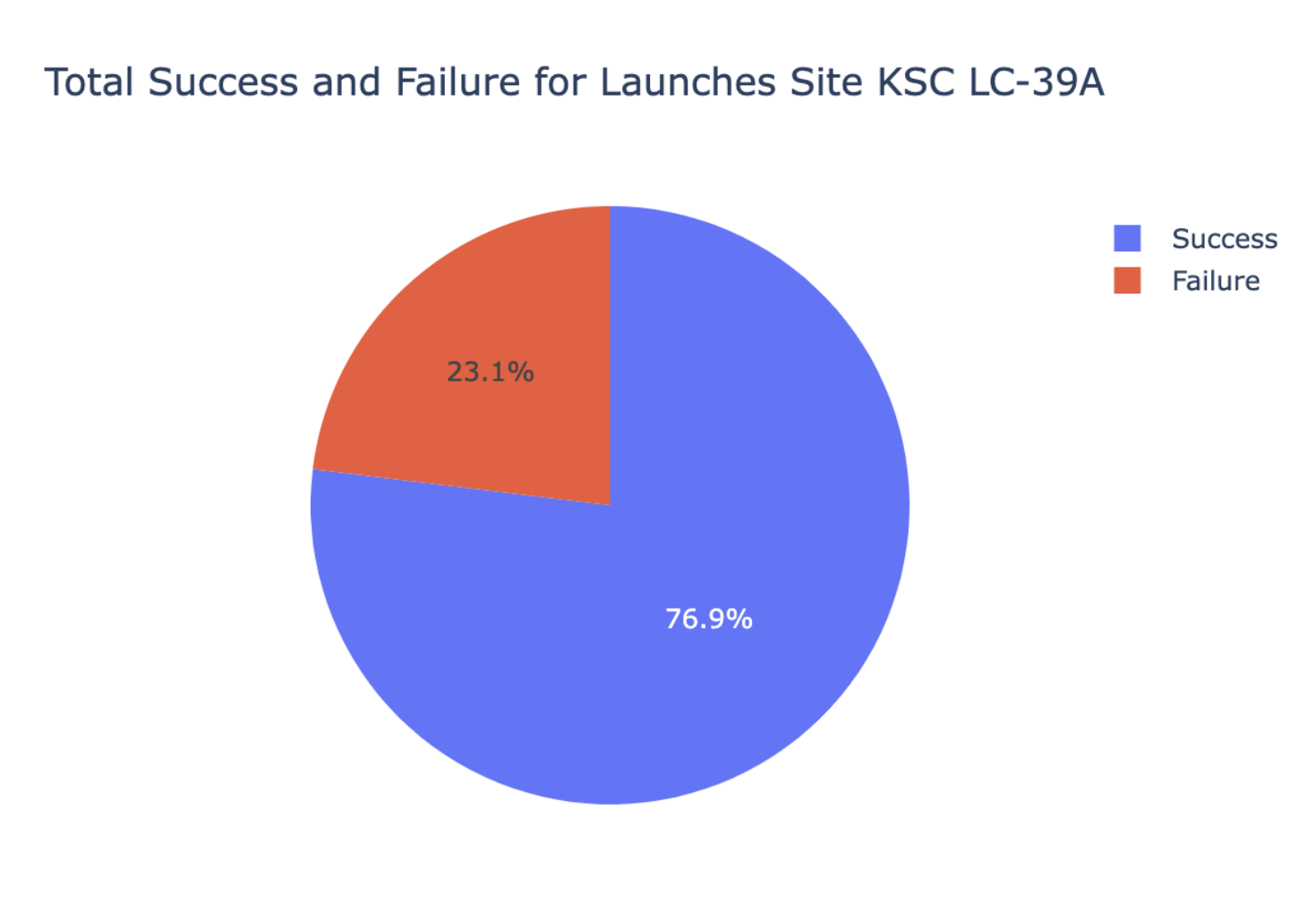

Section 4
Build a Dashboard with Plotly Dash

Launch Success counts per launch site



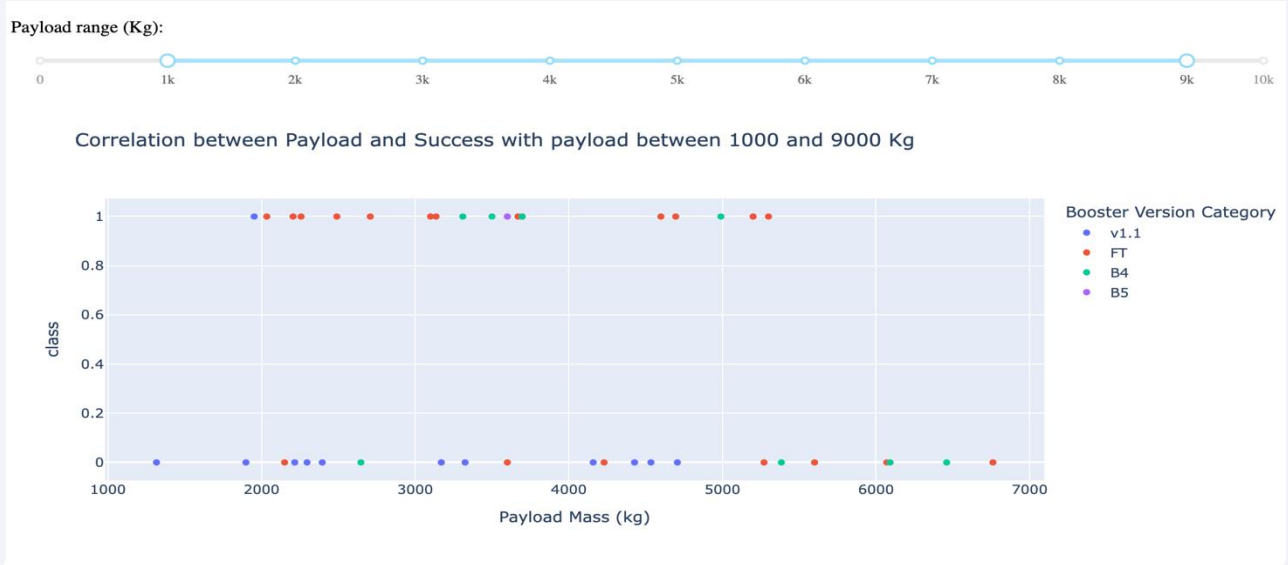
The pie chart shows that KSC LC-39 has the greatest share of successes (41.7%), followed by CCAFS LC-40 with 29.2 %

Outcome Ratio for KSC LC-39 Launch Site



The KSC LC-39 has 76.9% success in all rocket launches

Payload vs Launch Outcomes



Payload range slider allows users to select a range of payload masses (in kg) to filter the data displayed in the scatter plot.

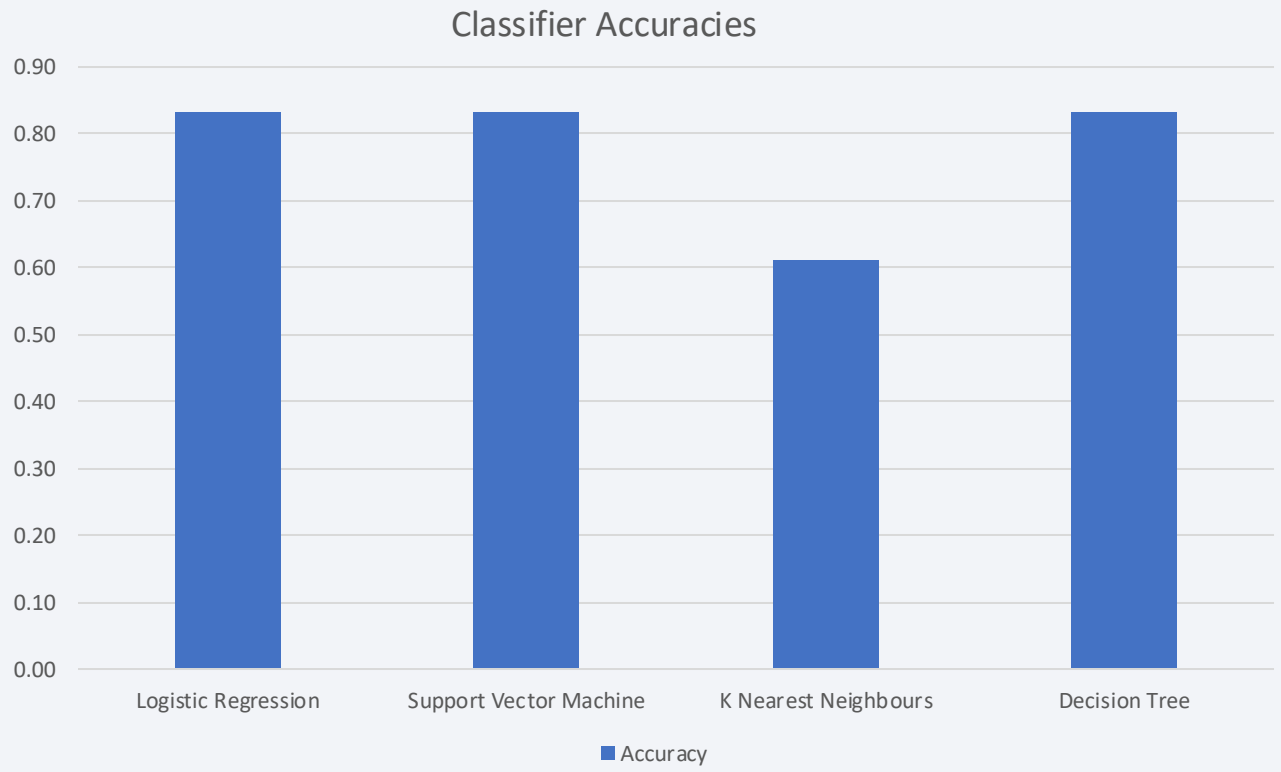
The plots show data for different payload ranges, such as 3000-6000 kg and 1000-9000 kg

The success and failure rate can be observed across all payload ranges

Section 5

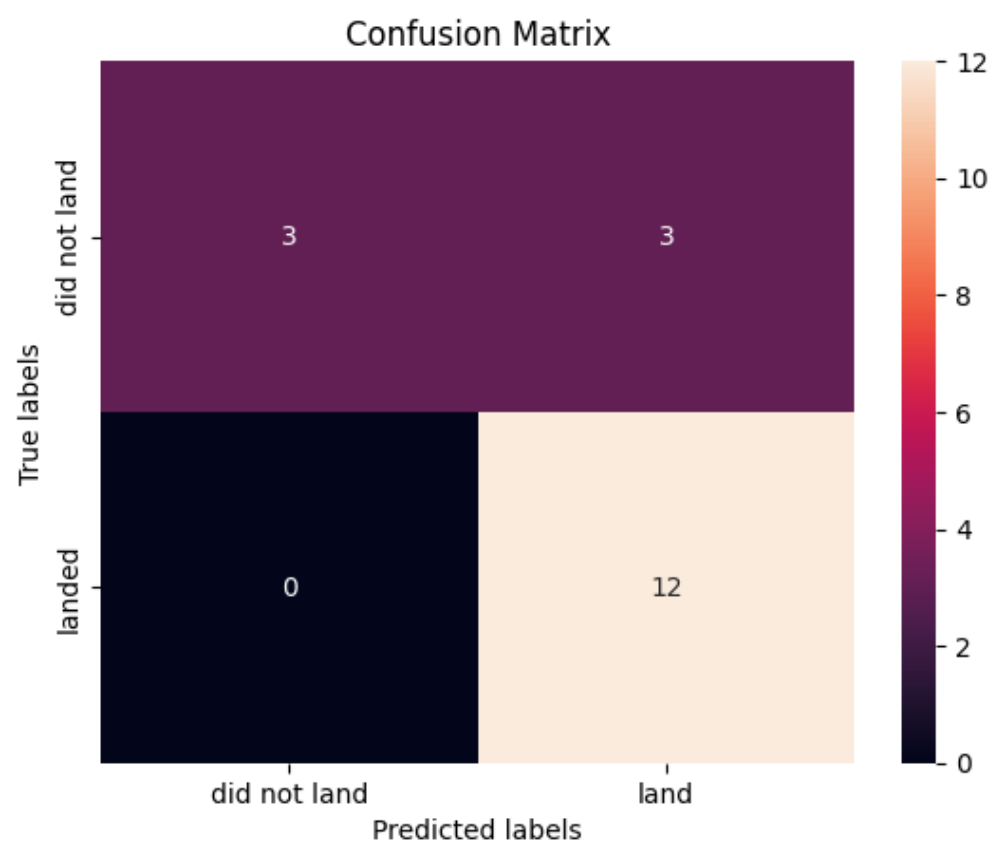
Predictive Analysis (Classification)

Classification Accuracy



- Logistic Regression, Support Vector Machine, and Decision Tree had the best accuracy of 83.33%
- K-Nearest Neighbors classifier had the least accuracy with 61 %

Confusion Matrix



- True Positives (TP) : The model correctly predicted 12 instances where the first stage landed.
- True Negatives (TN) : The model correctly predicted 3 instances where the first stage did not land.
- False Positives (FP): The model incorrectly predicted 3 instances as landed when they did not land (Type I error).
- False Negatives (FN): The model did not have any instances where it failed to predict a landing when it actually landed (Type II error).
- Overall, the model performs well with an accuracy of approximately 83.3%, but there is room for improvement in reducing false positives.

Conclusions

- SpaceX's Falcon 9 rocket offers launch services at a significantly lower cost ($62 million) compared to traditional providers ($165 million+), primarily due to the reusability of the rocket's first stage, provided it lands successfully.
- The project aimed to identify patterns and trends in landing outcomes through exploratory data analysis (EDA), interactive visualizations, and machine learning techniques. Key findings included:
    - Increased Landing Success Over Time: There has been a notable improvement in successful landings in recent years, especially from launch sites CCAFS SLC 40 and KSC LC 39A.
    - Influence of Target Orbits: Missions to orbits such as ES-L1, GEO, HEO, and SSO exhibited higher landing success rates. In contrast, missions targeting GTO and SO faced more challenges.
    - Predictive Modeling: A machine learning model was developed, achieving an 83% accuracy rate in predicting successful landings. Nonetheless, there is potential to enhance the model's performance, particularly in reducing false positive predictions.
- The find of will help SpaceX in strategizing launches for successful outcomes. Additional research is needed to improve the findings of this project.

Appendix

- The code, notebooks and data for the project can be download on GitHub at https://github.com/kennedyopokuasare/IBM_datascience/tree/main/captone_project

Thank you!